HAL
open science

Algorithms for nonnegative matrix factorization with the beta-divergence

Cédric Févotte, Jérôme Idier

To cite this version:

Cédric Févotte, Jérôme Idier. Algorithms for nonnegative matrix factorization with the beta-divergence. 2010. hal-00522289v2

HAL Id: hal-00522289

https://hal.science/hal-00522289v2

Preprint submitted on 8 Oct 2010 (v2), last revised 7 Mar 2011 (v4)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Algorithms for nonnegative matrix factorization with the β-divergence

Cédric Févotte[1] and Jérôme Idier[2]

[1] CNRS LTCI; Télécom ParisTech, France

Email : fevotte@telecom-paristech.fr

[2] CNRS IRCCyN; École Centrale de Nantes, France

October 1, 2010

Abstract

This paper describes algorithms for nonnegative matrix factorization (NMF) with the β-divergence (β-NMF). The β-divergence is a family of cost functions parametrized by a single shape parameter β that takes the Euclidean distance, the Kullback-Leibler divergence and the Itakura-Saito divergence as special cases ($\beta = 2, 1, 0$ respectively). The proposed algorithms are based on a surrogate *auxiliary function* (a local majorization of the criterion function). We first describe a *majorization-minimization* (MM) algorithm that leads to multiplicative updates, which differ from standard heuristic multiplicative updates by a β-dependent power exponent. The monotonicity of the heuristic algorithm can however be proven for $\beta \in (0, 1)$ using the proposed auxiliary function. Then we introduce the concept of *majorization-equalization* (ME) algorithm which produces updates that move along constant level sets of the auxiliary function and lead to larger steps than MM. Simulations on synthetic and real data illustrate the faster convergence of the ME approach. The paper also describes how the proposed algorithms can be extended to two common variants of NMF : penalized NMF (i.e., when a penalty function of the factors is added to the criterion function) and convex-NMF (when the dictionary is assumed to belong to a known subspace).

Keywords: Nonnegative matrix factorization (NMF), β-divergence, multiplicative algorithms, majorization-minimization (MM), majorization-equalization (ME).

1 Introduction

Given a data matrix $\mathbf{V}$ of dimensions $F \times N$ with nonnegative entries, NMF is the problem of finding a factorization

$$\mathbf{V} \approx \mathbf{W}\mathbf{H} \tag{1}$$

where $\mathbf{W}$ and $\mathbf{H}$ are nonnegative matrices of dimensions $F \times K$ and $K \times N$, respectively. K is usually chosen such that $F\,K + K\,N \ll F\,N$, hence reducing the data dimension. The factorization is in general only approximate, so that the terms "approximate nonnegative matrix factorization" or "nonnegative matrix approximation" also appear in the literature. NMF has been used for various problems in diverse fields. To cite a few, let us mention the problems of learning parts of faces and semantic features of text (Lee and Seung, 1999), polyphonic music transcription (Smaragdis and Brown, 2003), object characterization by reflectance spectra analysis (Berry et al., 2007), portfolio diversification (Drakakis et al., 2007), text mining (Xu et al., 2003), DNA gene expression analysis (Brunet et al., 2004), clustering of protein interactions (Greene et al., 2008), etc. The factorization (1) is usually sought after through the minimization problem

$$\min_{\mathbf{W},\mathbf{H}} D(\mathbf{V}|\mathbf{W}\mathbf{H}) \text{ subject to } \mathbf{W} \geq 0, \mathbf{H} \geq 0 \tag{2}$$

where the notation $\mathbf{A} \geq 0$ expresses nonnegativity of the entries of matrix $\mathbf{A}$ (and not semidefinite positiveness), and where $D(\mathbf{V}|\mathbf{W}\mathbf{H})$ is a separable measure of fit such that

$$D(\mathbf{V}|\mathbf{W}\mathbf{H}) = \sum_{f=1}^{F}\sum_{n=1}^{N} d([\mathbf{V}]_{fn}|[\mathbf{W}\mathbf{H}]_{fn}) \tag{3}$$

where $d(x|y)$ is a scalar cost function. What we intend by "cost function" is a positive function of $y \in \mathbb{R}_+$ given $x \in \mathbb{R}_+$, with a single minimum for $x = y$.

A popular cost function in NMF is the β-divergence $d_\beta(x|y)$ of Basu et al. (1998); Eguchi and Kano (2001); Cichocki et al. (2006), defined rigorously in Section 2.1. In essence, it is a parameterized cost function with a single parameter β, which takes the Euclidean distance, the generalized Kullback-Leibler (KL) divergence and the Itakura-Saito (IS) divergence as special cases ($\beta = 2, 1$ and 0, respectively). NMF with the β-divergence has been widely used in music signal processing in particular, for transcription and source separation (O'Grady, 2007; O'Grady and Pearlmutter, 2008; FitzGerald et al., 2009; Bertin et al., 2009; Févotte et al., 2009; Vincent et al., 2010; Dessein et al., 2010). In these work the nonnegative data matrix $\mathbf{V}$ is a spectrogram which is decomposed into elementary spectra with NMF. The parameter β can be tuned so as to optimize transcription or separation accuracy on training data. While popular in music signal processing, NMF with the β-divergence (shortened as "β-NMF" in the rest of the paper) can be of interest to any field.

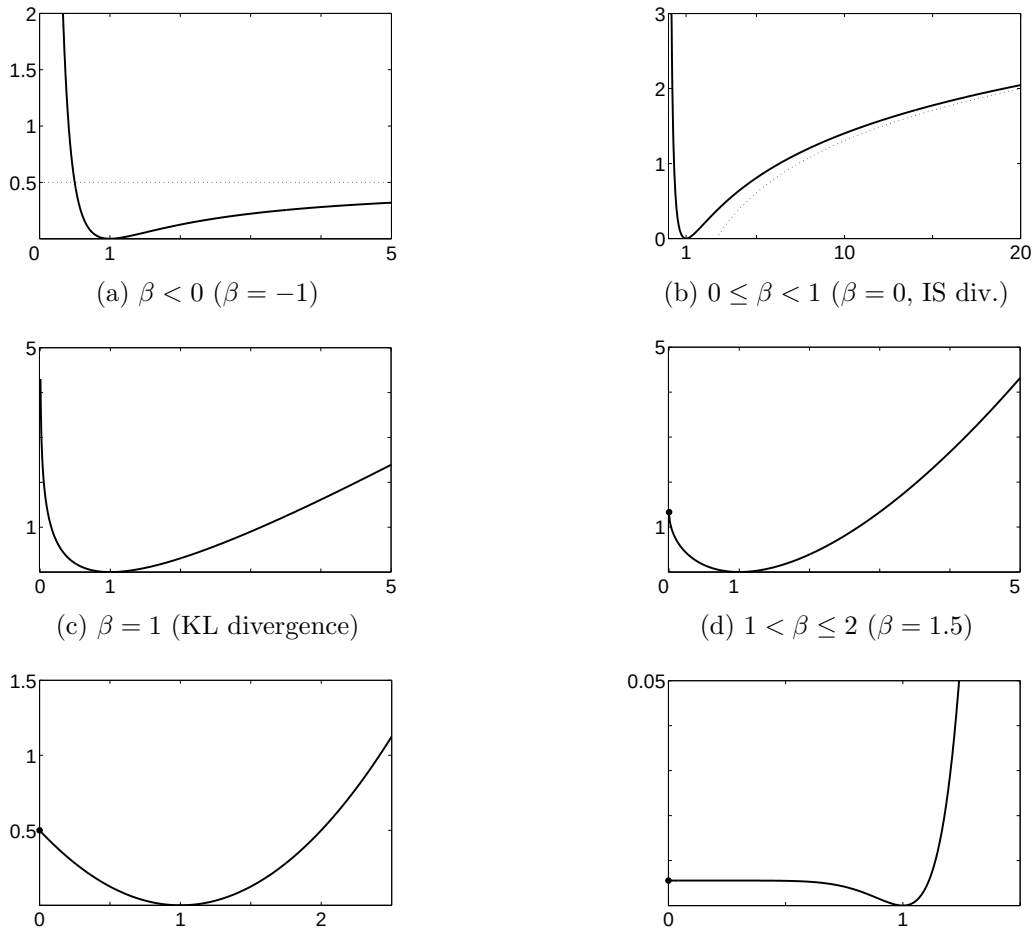
The NMF algorithm used in the latter papers is presented as a gradient-descent algorithm where the step size is set adaptatively and chosen such that the updates are multiplicative, as originally described in (Cichocki et al., 2006). The same algorithm can be derived from the following heuristic, proposed by Févotte et al. (2009). Let θ be a coefficient of $\mathbf{W}$ or $\mathbf{H}$. As will be seen later, when using the β-divergence the derivative $\nabla_\theta D(\theta)$ of the criterion $D(\mathbf{V}|\mathbf{WH})$ with respect to (w.r.t) θ can be expressed as the difference of two nonnegative functions such that $\nabla_\theta D(\theta) = \nabla_\theta^+ D(\theta) - \nabla_\theta^- D(\theta)$. Then, a heuristic multiplicative algorithm simply writes

$$\theta \leftarrow \theta . \frac{\nabla_\theta^- D(\theta)}{\nabla_\theta^+ D(\theta)} \tag{4}$$

which ensures nonnegativity of the parameter updates, provided initialization with a nonnegative value. A fixed point $\theta^\star$ of the algorithm implies either $\nabla_\theta D(\theta^\star) = 0$ or $\theta^\star = 0$. Monotonicity of this algorithm has been proven by (Kompass, 2007) for the specific range of values of β for which the divergence $d_\beta(x|y)$ is convex w.r.t y (i.e., $\beta \in [1, 2]$, see Section 2.1). The proof is based on a *majorization-minimization* (MM) procedure: an *auxiliary function* is built and iteratively minimized for each column of $\mathbf{H}$ (given $\mathbf{W}$) and each row of $\mathbf{W}$ (given $\mathbf{H}$). The auxiliary function is built using Jensen's inequality, thanks to convexity of the cost for $\beta \in [1, 2]$. However, it was observed in practice that the multiplicative algorithm (4) is still monotone (i.e., decreases the criterion function at each iteration) for values of β out of the "convexity" interval $[1, 2]$, though no proof is to avail.

This paper studies three descent algorithms for β-NMF, based on an auxiliary function that unifies existing auxiliary functions for the Euclidean distance and KL divergence (De Pierro, 1993; Lee and Seung, 2001), Kompass' divergence (Kompass, 2007) and the IS divergence (Cao et al., 1999). This auxiliary function was also recently proposed independently by Nakano et al. (2010). The construction of the auxiliary function relies on the decomposition of the criterion function into its convex and concave parts, following the approach of Cao et al. (1999) for the IS divergence. An auxiliary function to the convex part is constructed using Jensen's inequality while the concave part is locally majorized by its tangent. It is shown that MM algorithms based on the latter auxiliary function yield multiplicative updates that coincide with the heuristic described by Eq. (4) for $\beta \in [1, 2]$, but differ from a β-dependent power exponent when $\beta \notin [1, 2]$, a result also obtained by Nakano et al. (2010). Additionally, we show that the monotonicity of the heuristic algorithm can however be proven for $\beta \in (0, 1)$, using the proposed auxiliary function (it is shown to produce a descent algorithm though it does not fully minimize the auxiliary function). Then we introduce the concept of *maximization-equalization* (ME) algorithm which produces updates that move along constant level sets of the auxiliary and leads to larger steps than MM. This is akin to *overrelaxation* and is shown experimentally to produce faster convergence. Finally we show how the described MM, ME and heuristic algorithms can be extended to two common variants of NMF : penalized NMF (i.e., when a penalty function of $\mathbf{W}$ or $\mathbf{H}$ is added to the criterion function) and "convex"-NMF (when the dictionary is assumed to belong to a known subspace, as proposed by Ding et al. (2010)).

The paper is organized as follows. Section 2 defines and discusses the β-divergence, and then exposes in details the optimization task addressed in this paper. Section 3 recalls the concept of auxiliary function and then introduces a general auxiliary function for the β-NMF problem. Section 4 describes algorithms based on the proposed auxiliary function, namely MM and ME algorithms, and describe how they relate to the heuristic update (4). Section 5 reports simulations and convergence behaviors on synthetic and real data. Section 6 describes extensions of the proposed algorithms to penalized and convex- NMF. Section 7 concludes and discusses open questions.



(a) $\beta < 0$ ($\beta = -1$)

(b) $0 \leq \beta < 1$ ($\beta = 0$, IS div.)

(c) $\beta = 1$ (KL divergence)

(d) $1 < \beta \leq 2$ ($\beta = 1.5$)

(e) $\beta = 2$ (Euclidean distance)

(f) $\beta > 2$ ($\beta = 10$)

Figure 1: β-divergence $d_\beta(x|y)$ as a function of y (with $x = 1$). Subfigures (b), (c) and (e) illustrate Euclidean, KL and IS costs respectively. (Figure reproduced from (Bertin et al., 2009).)

2 Preliminaries

In this section we present the β-divergence and more precisely specify the task that is addressed in this paper. A detailed exposition of the β-divergence can be found in (Cichocki and Amari, 2010).

2.1 Definition of the β-divergence

The β-divergence was introduced by Basu et al. (1998) and Eguchi and Kano (2001) and can be defined as

$$d_\beta(x|y) \stackrel{\text{def}}{=} \begin{cases} \frac{1}{\beta\,(\beta-1)}\left(x^\beta + (\beta-1)\,y^\beta - \beta\,x\,y^{\beta-1}\right) & \beta \in \mathbb{R}\backslash\{0,1\} \\ x\,(\log x - \log y) + (y - x) & \beta = 1 \\ \frac{x}{y} - \log\frac{x}{y} - 1 & \beta = 0 \end{cases} \tag{5}$$

Basu et al. (1998) and Eguchi and Kano (2001) assume $\beta > 1$, but the definition domain can be extended to $\beta \in \mathbb{R}$, as suggested by Cichocki et al. (2006), which is the definition domain that is considered in this paper. The β-divergence can be shown continuous in β by using the identity $\lim_{\beta\to 0}(x^\beta - y^\beta)/\beta = \log(x/y)$. The β-divergence was considered for NMF by Cichocki et al. (2006) and also coincides up to a factor $1/\beta$ with the "generalized divergence" of Kompass (2007) which, in the context of NMF as well, was separately constructed so as to interpolate between the KL divergence ($\beta = 1$) and the Euclidean distance ($\beta = 2$). It takes the IS divergence as a special limit case, for $\beta = 0$. The β-divergence is plotted for various values of β on Fig. 1

The first and second derivative of $d_\beta(x|y)$ w.r.t y are also continuous in β, and write

$$d'_\beta(x|y) = y^{\beta-2}\,(y - x), \tag{6}$$

$$d''_\beta(x|y) = y^{\beta-3}\,\left[(\beta-1)y - (\beta-2)x\right]. \tag{7}$$

	$\breve{d}(x\|y)$	$\breve{d}'(x\|y)$	$\widehat{d}(x\|y)$	$\widehat{d}'(x\|y)$	$\bar{d}(x)$
$\beta < 1$ and $\beta \neq 0$	$-\frac{1}{\beta-1} x\, y^{\beta-1}$	$-x\, y^{\beta-2}$	$\frac{1}{\beta} y^{\beta}$	$y^{\beta-1}$	$\frac{1}{\beta(\beta-1)} x^{\beta}$
$\beta = 0$	$x\, y^{-1}$	$-x\, y^{-2}$	$\log y$	y^{-1}	$x(\log x - 1)$
$1 \leq \beta \leq 2$	$d_\beta(x\|y)$	$d'_\beta(x\|y)$	0	0	0
$\beta > 2$	$\frac{1}{\beta} y^{\beta}$	$y^{\beta-1}$	$-\frac{1}{\beta-1} x\, y^{\beta-1}$	$-x\, y^{\beta-2}$	$\frac{1}{\beta(\beta-1)} x^{\beta}$

Table 1: Example of differentiable convex-concave-constant decomposition of the β-divergence under the form (8).

The derivative shows that $d_\beta(x|y)$, as a function of y, has a single minimum in $y = x$ and that it increases with $|y-x|$, justifying its relevance as a measure of fit. The second derivative shows that the β-divergence is convex w.r.t y for $\beta \in [1, 2]$. Outside this interval the divergence can always be expressed as the sum of a convex, concave and constant part, such that

$$d_\beta(x|y) = \breve{d}(x|y) + \widehat{d}(x|y) + \bar{d}(x) \tag{8}$$

where $\breve{d}(x|y)$ is a convex function of y, $\widehat{d}(x|y)$ is a concave function of y and $\bar{d}(x)$ is a constant of y. The decomposition is not unique, since constant or linear terms (w.r.t y) are both convex and concave, or, less trivially, since any convex term can be added to $\breve{d}(x|y)$ while subtracted from $\widehat{d}(x|y)$. In the following we will use the "natural conventions" given in Table 1.

As noted by Févotte et al. (2009), a noteworthy property of the β-divergence is its behavior w.r.t to scale, as the following equation holds for any value of β :

$$d_\beta(\gamma\, x|\gamma\, y) = \gamma^\beta\; d_\beta(x|y). \tag{9}$$

It implies that factorizations obtained with $\beta > 0$ (such as with the Euclidean distance or the KL divergence) will rely more heavily on the largest coefficients and less precision is to be expected in the estimation of the low-power components, and conversely factorizations obtained with $\beta < 0$ will rely more heavily on smallest coefficients. The IS divergence ($\beta = 0$) is scale-invariant (i.e., $d_{IS}(\gamma\, x|\gamma\, y) = d_{IS}(x|y)$), and is the only one of the β-divergence family to possess this property. Factorizations with small values of β are relevant to decomposition of audio spectra, which typically exhibit exponential power decrease along frequency f and also usually comprise low-power transient components such as note attacks together with higher power components such as tonal parts of sustained notes. For example, Févotte et al. (2009) present the results of the decomposition of a piano power spectrogram with IS-NMF and show that components corresponding to very low residual noise and hammer hits on the strings are extracted with great accuracy, while these components are either ignored or severely degraded when using Euclidean or KL divergences. Similarly, the value $\beta = 0.5$ is advocated in (FitzGerald et al., 2009) and has been shown to give optimal results for music transcription based on NMF of the magnitude spectrogram in (Vincent et al., 2010).

It has also been noted by Févotte and Cemgil (2009) that the β-divergence belongs to the family of Bregman divergences, with potential function $\phi(y) = y^\beta/(\beta(\beta-1))$. NMF with Bregman divergences has been considered by Dhillon and Sra (2005), where the lack of results about the monotonicity of multiplicative algorithms in general has been noted.[1] This paper intends to fill this gap for the specific case of β-divergence.

2.2 Task

Core optimization problem As to our best knowledge all algorithms in the literature, the NMF algorithms we describe in this paper sequentially update $\mathbf{H}$ given $\mathbf{W}$ and then $\mathbf{W}$ given $\mathbf{H}$. These two steps are essentially the same, by symmetry of the factorization ($\mathbf{V} \approx \mathbf{WH}$ is equivalent to $\mathbf{V}^T \approx \mathbf{H}^T\mathbf{W}^T$ and the roles of $\mathbf{W}$ and $\mathbf{H}$ are simply exchanged), and because we are not making any assumption on the relative values of F and N. Hence, we may concentrate on solving the following subproblem

$$\min_{\mathbf{H}} C(\mathbf{H}) \stackrel{\text{def}}{=} D(\mathbf{V}|\mathbf{WH}) \text{ subject to } \mathbf{H} \geq 0 \tag{10}$$

with fixed $\mathbf{W}$ and where in the rest of the paper $D(\mathbf{V}|\mathbf{WH})$ is as of Eq. (3) with $d(x|y) = d_\beta(x|y)$. The criterion function $C(\mathbf{H})$ separates into $\sum_n D(\mathbf{v}_n|\mathbf{W}\mathbf{h}_n)$, where $\mathbf{v}_n$ and $\mathbf{h}_n$ are the nth row of $\mathbf{V}$ and $\mathbf{H}$,

[1] More precisely, Dhillon and Sra (2005) give proofs of monotonicity for the "reverse" problem of minimizing $D(\mathbf{WH}|\mathbf{V})$ instead of $D(\mathbf{V}|\mathbf{WH})$, while pointing that monotonicity of multiplicative algorithms based on the heuristic (4) for the latter problem is however observed in practice.

respectively, so that we are essentially left with solving the problem

$$\min_{\mathbf{h}} C(\mathbf{h}) = D(\mathbf{v}|\mathbf{W}\mathbf{h}) \text{ subject to } \mathbf{h} \geq 0 \tag{11}$$

where $\mathbf{v} \in \mathbb{R}_+^F$, $\mathbf{W} \in \mathbb{R}_+^{F\times K}$ and $\mathbf{h} \in \mathbb{R}_+^K$.

KKT necessary conditions An admissible solution $\mathbf{h}^\star$ to problem (11) must satisfy the Karush-Kuhn-Tucker (KKT) first order optimality conditions, which write

$$\nabla_{\mathbf{h}} C(\mathbf{h}^\star).\mathbf{h}^\star = 0 \tag{12}$$

$$\nabla_{\mathbf{h}} C(\mathbf{h}^\star) \geq 0 \tag{13}$$

$$\mathbf{h}^\star \geq 0 \tag{14}$$

where the dot notation '.' denotes entrywise operations (here term-to-term multiplication) and $\nabla_{\mathbf{h}} C(\mathbf{h})$ denotes the gradient of $C(\mathbf{h})$, given by

$$\nabla_{\mathbf{h}} C(\mathbf{h}) = \mathbf{W}^T \left[d'(v_f|[\mathbf{W}\mathbf{h}]_f)\right]_f \tag{15}$$

$$= \mathbf{W}^T[(\mathbf{W}\mathbf{h})^{\cdot(\beta-2)}(\mathbf{W}\mathbf{h} - \mathbf{v})] \tag{16}$$

where the notation $[x_f]_f$ refers to the column vector $[x_1, \ldots, x_F]^T$. The KKT conditions (12)-(14) can be summarized as

$$\min\{\mathbf{h}^\star, \nabla_{\mathbf{h}} C(\mathbf{h}^\star)\} = \mathbf{0}_K \tag{17}$$

where the "min" operator is entrywise and $\mathbf{0}_K$ is a null vector of dimension K.

Algorithms In the following, we will say that an algorithm is *monotone* if it produces a sequence of iterates $\{\mathbf{h}^{(i)}\}_{i\geq 0}$, $i \geq 0$ such that $C(\mathbf{h}^{(i+1)}) \leq C(\mathbf{h}^{(i)})$. An algorithm is said *convergent* if it produces a sequence of iterates $\{\mathbf{h}^{(i)}\}_{i\geq 0}$ which converges to a limit point $\mathbf{h}^\star$ satisfying the KKT conditions (12)-(14). Monotonicity does not imply convergence in general, nor is monotonicity necessary to convergence.

3 An auxiliary function for β-NMF

In this section we properly define the concept of auxiliary function and then exhibit a separable auxiliary function for the β-NMF problem.

3.1 Definition of auxiliary function

Definition 1 (Auxiliary function). The $\mathbb{R}_+^K \times \mathbb{R}_+^K \to \mathbb{R}_+$ mapping $G(\mathbf{h}|\tilde{\mathbf{h}})$ is said to be an *auxiliary function* to $C(\mathbf{h})$ if and only if

- $\forall \mathbf{h} \in \mathbb{R}_+^K$, $C(\mathbf{h}) = G(\mathbf{h}|\mathbf{h})$
- $\forall (\mathbf{h}, \tilde{\mathbf{h}}) \in \mathbb{R}_+^K \times \mathbb{R}_+^K$, $C(\mathbf{h}) \leq G(\mathbf{h}|\tilde{\mathbf{h}})$

In other words an auxiliary function $G(\mathbf{h}|\tilde{\mathbf{h}})$ is a *majorizing function* or *upper bound* of $C(\mathbf{h})$ which is tight for $\mathbf{h} = \tilde{\mathbf{h}}$. The optimization of $C(\mathbf{h})$ can be replaced by iterative optimization of $G(\mathbf{h}|\tilde{\mathbf{h}})$. Indeed, any iterate $\mathbf{h}^{(i+1)}$ satisfying

$$G(\mathbf{h}^{(i+1)}|\mathbf{h}^{(i)}) \leq G(\mathbf{h}^{(i)}|\mathbf{h}^{(i)}) \tag{18}$$

satisfies $C(\mathbf{h}^{(i+1)}) \leq C(\mathbf{h}^{(i)})$, because we have

$$C(\mathbf{h}^{(i+1)}) \leq G(\mathbf{h}^{(i+1)}|\mathbf{h}^{(i)}) \leq G(\mathbf{h}^{(i)}|\mathbf{h}^{(i)}) = C(\mathbf{h}^{(i)}). \tag{19}$$

The iterate $\mathbf{h}^{(i+1)}$ is typically chosen as

$$\mathbf{h}^{(i+1)} = \arg\min_{\mathbf{h}\geq 0} \; G(\mathbf{h}|\mathbf{h}^{(i)}) \tag{20}$$

which forms the basis of *maximization-minimization* (MM) algorithms (Hunter and Lange, 2004). However, any other iterate $\mathbf{h}^{(i+1)}$ satisfying (18) produces a monotone algorithm. As such, Figure 2 illustrates the three updates strategies that will be developed in this paper.

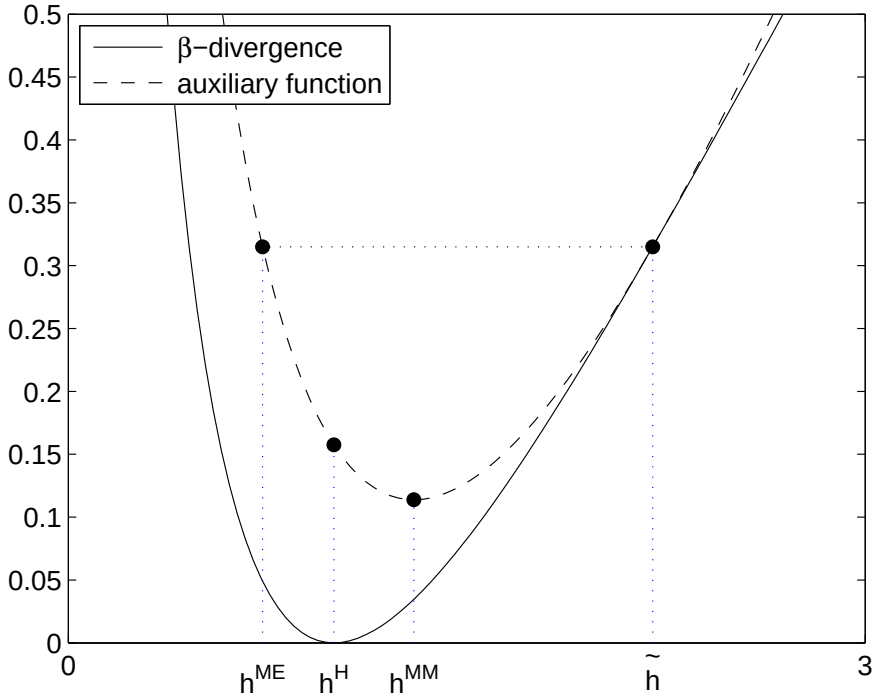


Figure 2: The β-divergence $d_\beta(x|y)$ for $\beta = 0.5$ (with $x = 1$) and its auxiliary function in dimension one (with $\tilde{h} = 2.2$). The MM update h^{MM} corresponds to the minimum of the auxiliary function, see Section 4.1. The heuristic update h^{H} is discussed in Section 4.2 (the heuristic update minimizes the criterion function in the simple one-dimensional case but this is not true in larger dimensions). The ME update h^{ME} consists in selecting the next update "beyond the valley" defined by the auxiliary function, from the current solution $\tilde{h}$, see Section 4.3.

3.2 Separable auxiliary function for β-NMF

In this section we construct an auxiliary function to $C(\mathbf{h})$ for the specific case of the β-divergence. Our approach follows the one of Cao et al. (1999) for IS divergence, and consists of majorizing the convex part of the criterion using Jensen's inequality and majorizing the concave part by its tangent, as detailed in the proof of the following theorem. Here and henceforth, we denote $\mathbf{W}\tilde{\mathbf{h}}$ by $\tilde{\mathbf{v}}$, with entries $[\mathbf{W}\tilde{\mathbf{h}}]_f = \tilde{v}_f$.

Theorem 1 (Auxiliary function for β-NMF). Let $\tilde{\mathbf{h}}$ be such that

(i) $\forall f, \tilde{v}_f > 0$

(ii) $\forall k, \tilde{h}_k > 0$

Then, the function $G(\mathbf{h}|\tilde{\mathbf{h}})$ defined by

$$G(\mathbf{h}|\tilde{\mathbf{h}}) = \sum_f \left[\sum_k \frac{w_{fk}\tilde{h}_k}{\tilde{v}_f} \breve{d}\left(v_f | \tilde{v}_f \frac{h_k}{\tilde{h}_k} \right) \right] + \left[\widehat{d}'(v_f|\tilde{v}_f) \sum_k w_{fk}(h_k - \tilde{h}_k) + \widehat{d}(v_f|\tilde{v}_f) \right] + \bar{d}(v_f) \tag{21}$$

is an auxiliary function to $C(\mathbf{h}) = \sum_f d(v_f|[\mathbf{W}\mathbf{h}]_f)$, where $\breve{d}(x|y) + \widehat{d}(x|y) + \bar{d}(x)$ is any differentiable convex-concave-constant decomposition of the β-divergence, such as the one defined in Table 1.

Proof. The condition $G(\mathbf{h}|\mathbf{h}) = C(\mathbf{h})$ is trivially met. The criterion $C(\mathbf{h})$ may be written as

$$C(\mathbf{h}) = \sum_f C_f(\mathbf{h}) \tag{22}$$

where $C_f(\mathbf{h}) \stackrel{\text{def}}{=} d(v_f|[\mathbf{W}\mathbf{h}]_f)$. We prove $C(\mathbf{h}) \leq G(\mathbf{h}|\tilde{\mathbf{h}})$ by constructing an auxiliary function to each part $C_f(\mathbf{h})$ of the criterion, and more precisely by treating the convex and concave part separately. Let us define $\breve{C}_f(\mathbf{h}) \stackrel{\text{def}}{=} \breve{d}(v_f|[\mathbf{W}\mathbf{h}]_f)$ and $\widehat{C}_f(\mathbf{h}) \stackrel{\text{def}}{=} \widehat{d}(v_f|[\mathbf{W}\mathbf{h}]_f)$, so that we can write

$$C_f(\mathbf{h}) = \breve{C}_f(\mathbf{h}) + \widehat{C}_f(\mathbf{h}) + \bar{d}(v_f). \tag{23}$$

Convex part: We first prove that

$$\breve{G}_f(\mathbf{h}|\tilde{\mathbf{h}}) = \sum_k \frac{w_{fk}\tilde{h}_k}{\tilde{v}_f}\ \breve{d}\left(v_f|\tilde{v}_f\frac{h_k}{\tilde{h}_k}\right) \tag{24}$$

is an auxiliary function to $\breve{C}_f(\mathbf{h})$. The condition $\breve{G}_f(\mathbf{h}|\mathbf{h}) = \breve{C}_f(\mathbf{h})$ is trivially met. The condition $\breve{G}_f(\mathbf{h}|\tilde{\mathbf{h}}) \geq \breve{C}_f(\tilde{\mathbf{h}})$ is proven as follows. Let $\mathcal{K}$ be the set of indices k such that $w_{fk} \neq 0$. Define $\forall k \in \mathcal{K}$,

$$\tilde{\lambda}_{fk} = \frac{w_{fk}\tilde{h}_k}{\tilde{v}_f} = \frac{w_{fk}\tilde{h}_k}{\sum_{\ell\in\mathcal{K}} w_{f\ell}\tilde{h}_{\ell n}}. \tag{25}$$

We have $\sum_{k\in\mathcal{K}} \tilde{\lambda}_{fk} = 1$ and

$$\begin{aligned}
\breve{G}_f(\mathbf{h}|\tilde{\mathbf{h}}) &= \sum_{k\in\mathcal{K}} \tilde{\lambda}_{fk}\ \breve{d}\left(v_f|\frac{w_{fk}h_k}{\tilde{\lambda}_{fk}}\right) &(26)\\
&\geq \breve{d}\left(v_f|\sum_{k\in\mathcal{K}} \tilde{\lambda}_{fk}\frac{w_{fk}h_k}{\tilde{\lambda}_{fk}}\right) &(27)\\
&= \breve{d}\left(v_f|\sum_{k=1}^{K} w_{fk}h_k\right) &(28)\\
&= \breve{C}_f(\mathbf{h}) &(29)
\end{aligned}$$

where we used Jensen's inequality, by convexity of $\breve{d}(x|y)$.

Concave part: An auxiliary function $\widehat{G}_f(\mathbf{h}|\tilde{\mathbf{h}})$ to the concave part $\widehat{C}_f(\mathbf{h})$ can be taken as the first order Taylor approximation to $\widehat{C}_f(\mathbf{h})$ in the vicinity of $\tilde{\mathbf{h}}$, i.e.,

$$\widehat{G}_f(\mathbf{h}|\tilde{\mathbf{h}}) = \widehat{C}_f(\tilde{\mathbf{h}}) + \nabla^T\widehat{C}_f(\tilde{\mathbf{h}})\ (\mathbf{h} - \tilde{\mathbf{h}}). \tag{30}$$

The function satisfies $\widehat{G}_f(\mathbf{h}|\mathbf{h}) = \widehat{C}_f(\mathbf{h})$ by construction and $\widehat{G}_f(\mathbf{h}|\tilde{\mathbf{h}}) \geq \widehat{C}_f(\mathbf{h})$ by concavity of $\widehat{C}_f(\mathbf{h})$, using the property that the tangent to any point is an upper bound of a concave function.[2] Using

$$\nabla_{h_k}\widehat{C}_f(\mathbf{h}) = w_{fk}\widehat{d}'(v_f|[\mathbf{Wh}]_f) \tag{31}$$

the explicit form for $\widehat{G}_f(\mathbf{h}|\tilde{\mathbf{h}})$ is given by

$$\widehat{G}_f(\mathbf{h}|\tilde{\mathbf{h}}) = \widehat{d}(v_f|\tilde{v}_f) + \widehat{d}'(v_f|\tilde{v}_f)\sum_k w_{fk}(h_k - \tilde{h}_k). \tag{32}$$

In the end a suitable auxiliary function $G(\mathbf{h}|\tilde{\mathbf{h}})$ to $C(\mathbf{h})$ is obtained by summing up the auxiliary functions constructed for each individual part of the criterion, i.e.,

$$G(\mathbf{h}|\tilde{\mathbf{h}}) = \sum_f \left(\breve{G}_f(\mathbf{h}|\tilde{\mathbf{h}}) + \widehat{G}_f(\mathbf{h}|\tilde{\mathbf{h}}) + \bar{d}(v_f)\right) \tag{33}$$

which leads to Eq. (21). □

Properties of the auxiliary function $G(\mathbf{h}|\tilde{\mathbf{h}})$ is by construction separable in functions of the individual coefficients h_k of $\mathbf{h}$, which allows to decouple the optimization. It is convenient to rewrite the auxiliary function as such in order to derive some of the algorithms of Section (4). We may write

$$G(\mathbf{h}|\tilde{\mathbf{h}}) = \sum_k G_k(h_k|\tilde{\mathbf{h}}) + cst \tag{34}$$

where *cst* is a constant w.r.t $\mathbf{h}$ and

$$G_k(h_k|\tilde{\mathbf{h}}) \stackrel{\text{def}}{=} \tilde{h}_k\left[\sum_f \frac{w_{fk}}{\tilde{v}_f}\breve{d}\left(v_f|\tilde{v}_f\frac{h_k}{\tilde{h}_k}\right)\right] + h_k\left[\sum_f w_{fk}\widehat{d}'(v_f|\tilde{v}_f)\right]. \tag{35}$$

[2] $\widehat{C}_f(\mathbf{h}) = \widehat{d}(v_f|[\mathbf{Wh}]_f)$ is concave as the composition of a concave function and a linear function.

	$\beta < 1$	$1 \leq \beta \leq 2$	$\beta > 2$
$\gamma(\beta)$	$\frac{1}{2-\beta}$	1	$\frac{1}{\beta-1}$

Table 2: Exponent in the multiplicative updates given by the MM algorithm.

The gradient of the auxiliary function is given by

$$\nabla_{h_k} G(\mathbf{h}|\tilde{\mathbf{h}}) = \sum_f w_{fk} \left[\breve{d}' \left(v_f | \tilde{v}_f \frac{h_k}{\tilde{h}_k} \right) + \widehat{d}'(v_f|\tilde{v}_f) \right]. \tag{36}$$

Thanks the separability of the auxiliary function into its variables the Hessian matrix is diagonal with

$$\nabla^2_{h_k} G(\mathbf{h}|\tilde{\mathbf{h}}) = \sum_f \tilde{v}_f \frac{w_{fk}}{\tilde{h}_k} \breve{d}'' \left(v_f | \tilde{v}_f \frac{h_k}{\tilde{h}_k} \right). \tag{37}$$

By convexity of $\breve{d}(x|y)$ we have $\breve{d}''(x|y) \geq 0$ which implies positive definiteness of the Hessian matrix and hence convexity of the auxiliary function $G(\mathbf{h}|\tilde{\mathbf{h}})$ (convexity more simply derives from the fact that the auxiliary function is built as a sum of convex functions).

Connections with other works The construction of $G(\mathbf{h}|\tilde{\mathbf{h}})$ employs standard mathematical tools (Jensen's inequality, Taylor approximation) that are well known from the MM literature, see, e.g., (Hunter and Lange, 2004). For $\beta \in [1, 2]$, $G(\mathbf{h}|\tilde{\mathbf{h}})$ coincides with the auxiliary function built in (Kompass, 2007). This latter paper proposed itself a generalization of the auxiliary functions proposed in (Lee and Seung, 2001) for the Euclidean distance ($\beta = 2$) and the generalized KL divergence ($\beta = 1$). For $\beta = 0$ (IS divergence), $G(\mathbf{h}|\tilde{\mathbf{h}})$ coincides with the auxiliary function proposed in (Cao et al., 1999). It is worth recalling that in the algorithms proposed by Lee and Seung (2001) the update of $\mathbf{W}$ given $\mathbf{H}$ or $\mathbf{H}$ given $\mathbf{W}$ are instances of well known algorithms for image restoration (for which $\mathbf{W}$ acts as a fixed, known blurring matrix and $\mathbf{H}$ is a vectorized image to be reconstructed). These algorithms are the Iterative Space Reconstructing Algorithm (ISRA) (Daube-Witherspoon and Muehllehner, 1986) and the Richardson-Lucy (RL) algorithm (Richardson, 1972; Lucy, 1974), which perform nonnegative linear regression with the Euclidean distance and KL divergence, respectively. The ISRA and RL algorithms are shown to be MM algorithms by De Pierro (1993). Similarly, the algorithms proposed by Cao et al. (1999) for nonnegative linear regression with the IS divergence were designed in the image restoration setting.

4 Algorithms for β-NMF

In section we describe algorithms for β-NMF based on the auxiliary function constructed in the latter section. In the following $\tilde{\mathbf{h}}$ should be understood as the current iterate $\mathbf{h}^{(i)}$ and we are seeking to obtain $\mathbf{h}^{(i+1)}$ such that Eq. (18) is satisfied.

4.1 Maximization-Minimization (MM) algorithm

An MM algorithm can be derived by minimizing the auxiliary function $G(\mathbf{h}|\tilde{\mathbf{h}})$ w.r.t to $\mathbf{h}$. Given the convexity and the separability of the auxiliary function the optimum is obtained by cancelling the gradient given by Eq. (36). This is trivially done and leads to the following update:

$$h_k^{\mathrm{MM}} = \tilde{h}_k \left(\frac{\sum_f w_{fk}\, v_f\, \tilde{v}_f^{\beta-2}}{\sum_f w_{fk}\, \tilde{v}_f^{\beta-1}} \right)^{\gamma(\beta)} \tag{38}$$

where $\gamma(\beta)$ is given in Table 2. Note that $\gamma(\beta) \leq 1, \forall \beta$. As suggested in Section 1, the gradient of the criterion may be written as the difference of two nonnegative functions such that

$$\nabla_{h_k} C(\tilde{\mathbf{h}}) = \nabla^+_{h_k} C(\tilde{\mathbf{h}}) - \nabla^-_{h_k} C(\tilde{\mathbf{h}}) \tag{39}$$

$$\nabla^+_{h_k} C(\tilde{\mathbf{h}}) = \sum_f w_{fk}\, \tilde{v}_f^{\beta-1} \tag{40}$$

$$\nabla^-_{h_k} C(\tilde{\mathbf{h}}) = \sum_f w_{fk}\, v_f\, \tilde{v}_f^{\beta-2} \tag{41}$$

so that the update (38) can be rewritten in the more interpretable form

$$h_k^{\mathrm{MM}} = \tilde{h}_k \left(\frac{\nabla_{h_k}^- C(\tilde{\mathbf{h}})}{\nabla_{h_k}^+ C(\tilde{\mathbf{h}})} \right)^{\gamma(\beta)}. \tag{42}$$

The conclusion is thus that the MM algorithm leads to multiplicative updates, but they differ from the "usual ones" derived heuristically by Cichocki et al. (2006) through gradient descent with adaptative step or by Févotte et al. (2009) by splitting the gradient into two nonnegative functions as above. They differ by the exponent $\gamma(\beta)$ which is not equal to one for $\beta \notin [1, 2]$.

4.2 Heuristic algorithm

Very few mathematical results have been established for the heuristic update

$$h_k^{\mathrm{H}} = \tilde{h}_k \left(\frac{\sum_f w_{fk}\, v_f\, \tilde{v}_f^{\beta-2}}{\sum_f w_{fk}\, \tilde{v}_f^{\beta-1}} \right) \tag{43}$$

when β falls outside $[1, 2]$, i.e., when the β-divergence $d_\beta(x|y)$ is not convex. In such a case, the heuristic update can be erroneously interpreted as an MM algorithm by wrongly applying Jensen's inequality to $C(\mathbf{h})$. Yet, in the particular case $\beta = 0$, it holds that each heuristic update produces a decrease of $C(\mathbf{h})$ (Cao et al., 1999). In the present subsection, this result is extended to all values of β between 0 and 1.

Let us first introduce a scalar auxiliary function $g(y|\tilde{y}; x)$ as follows:

$$\forall y, \tilde{y}, x > 0, \quad g(y|\tilde{y}; x) = \breve{d}(x|y) + \widehat{d}(x|\tilde{y}) + (y - \tilde{y})\widehat{d}'(x|\tilde{y}) + \bar{d}(x) \tag{44}$$

where $\breve{d}(x|y)$, $\widehat{d}(x|y)$ and $\bar{d}(x|y)$ are defined in Table 1. By immediate application of Theorem 1 to the scalar case, $g(y|\tilde{y}; x)$ is an auxiliary function to $d(x|y)$. In particular, $g(\tilde{y}|\tilde{y}; x) = d(x|\tilde{y})$. Then, we have the following preliminary result.

Lemma 1. For all $\beta \in \mathbb{R}$,

$$G_k(h_k|\tilde{\mathbf{h}}) = \frac{1}{\tilde{h}_k^{\beta-1}} \left(\sum_f w_{fk} \tilde{v}_f^{\beta-1} \right) g(h_k|\tilde{h}_k; h_k^{\mathrm{H}}) + cst. \tag{45}$$

Proof. For each of the four possible expressions of $(\widehat{d}, \breve{d})$ given in Table 1, the validity of (45) can be checked straightforwardly by direct verification. □

As already mentioned in Section 3.1, the MM update (20) is not the only way of taking advantage of the auxiliary function $G(\mathbf{h}|\tilde{\mathbf{h}})$ to obtain a decrease of $C(\mathbf{h})$: any update satisfying (18) also ensures that $C(\mathbf{h})$ does not increase. This is a key remark to understand the behavior of the heuristic algorithm for $\beta \in (0, 1)$, given the following property.

Theorem 2. For all $\beta \in (0, 1)$, and all $\tilde{\mathbf{h}}$ such that Conditions (i)-(ii) of Theorem 1 hold, the heuristic algorithm produces nonincreasing values of $C(\mathbf{h})$, according to the following inequality:

$$G(\mathbf{h}^{\mathrm{H}}|\tilde{\mathbf{h}}) \leq G(\tilde{\mathbf{h}}|\tilde{\mathbf{h}}). \tag{46}$$

Proof. For all $\beta \in (0, 1)$, straightforward calculations yield

$$g(\tilde{y}|\tilde{y}; x) - g(x|\tilde{y}; x) = \breve{d}(x|\tilde{y}) - \breve{d}(x|x) - (x - \tilde{y})\widehat{d}'(x|\tilde{y}) \tag{47}$$

$$= \frac{1}{1-\beta}\, \tilde{y}^\beta (1 - \beta + \beta\theta - \theta^\beta) \tag{48}$$

where $\theta = x/\tilde{y}$. Since $f(\theta) = \theta^\beta$ is a concave function of θ, we have $f(\theta) \leq f(1) + (\theta - 1)f'(1)$, which also reads $\theta^\beta \leq 1 + (\theta - 1)\beta$. Hence, $g(\tilde{y}|\tilde{y}; x) - g(x|\tilde{y}; x) \geq 0$ for all x, $\tilde{y}$. The latter inequality implies $\forall k$, $g(h_k^{\mathrm{H}}|\tilde{h}_k, h_k^{\mathrm{H}}) \leq g(\tilde{h}_k|\tilde{h}_k, h_k^{\mathrm{H}})$, so that we have $G_k(h_k^{\mathrm{H}}|\tilde{\mathbf{h}}) \leq G_k(\tilde{h}_k|\tilde{\mathbf{h}})$ according to (45), which leads to the result by summation over k. □

Cao et al. (1999) show that inequality (46) becomes an equality in the case $\beta = 0$, so that each heuristic update yields $G(\mathbf{h}^{\mathrm{H}}|\tilde{\mathbf{h}}) = G(\tilde{\mathbf{h}}|\tilde{\mathbf{h}})$. In this particular case, the heuristic algorithm can be called a "majorization-equalization" algorithm, a class of algorithms described in next section. For values of β outside the range $[0, 2]$, the inequality (46) does not hold anymore.[3] Of course, this does not mean that the heuristic updates produce increasing values of $C(\mathbf{h})$. On the contrary, numerical simulations tend to indicate that they always produce nonincreasing values of $C(\mathbf{h})$, but proving this is still an open issue for $\beta \notin [0, 2]$. Compared to MM updates, heuristic updates produce larger or equal steps for all β, since it can trivially be shown that

$$\forall k, \quad |h_k^{\mathrm{H}} - \tilde{h}_k| \geq |h_k^{\mathrm{MM}} - \tilde{h}_k|. \tag{49}$$

For $\beta \notin [1, 2]$, numerical simulations indicate that the heuristic algorithm is faster than the MM algorithm (and we recall that the two algorithms coincide for $\beta \in [1, 2]$). Given (49), skipping from the latter to the former has an effect comparable to that of overrelaxation: on the average, stretching the steps allows to reduce their number to reach convergence.

In order to produce even larger steps for $\beta \in [0, 2]$, and yet nonincreasing values of $C(\mathbf{h})$, the following subsection explores the concept of majorization-equalization.

4.3 Majorization-Equalization (ME) algorithm

Let us introduce the general notion of ME update by the fact that the new iterate $\mathbf{h}^{\mathrm{ME}}$ fulfills

$$G(\mathbf{h}^{\mathrm{ME}}|\tilde{\mathbf{h}}) = G(\tilde{\mathbf{h}}|\tilde{\mathbf{h}}). \tag{50}$$

Eq. (50) actually defines a level set rather than a single point. Let us concentrate on the following more constrained and manageable condition, given the separability of $G(\mathbf{h}|\tilde{\mathbf{h}})$:

$$\forall k, \quad G_k(h_k^{\mathrm{ME}}|\tilde{\mathbf{h}}) = G_k(\tilde{h}_k|\tilde{\mathbf{h}}).$$

Given (45), this amounts to solve the following equation for y, for any $\tilde{y}, x > 0$:

$$g(y|\tilde{y}; x) = g(\tilde{y}|\tilde{y}; x). \tag{51}$$

Since $g(y|\tilde{y}; x)$ is strictly convex w.r.t y, (51) has not more than two solutions, one of them being $\tilde{y}$. By construction, the selection of the other solution (provided that it exists) will provide ME steps that are larger than MM updates, i.e.,

$$\forall k, \quad |h_k^{\mathrm{ME}} - \tilde{h}_k| \geq |h_k^{\mathrm{MM}} - \tilde{h}_k|, \tag{52}$$

as illustrated by Figure 2. To go further on the determination of this solution, a case-by-case analysis must be performed, depending on the range of β.

Case 1 : $\beta \in [0, 1)$ In that case we have

$$g(y|\tilde{y}; x) = \frac{1}{1-\beta} x\, y^{\beta-1} + y\tilde{y}^{\beta-1} + cst. \tag{53}$$

Let us remark that

$$\forall\, \tilde{y}, x > 0, \quad \lim_{y \to 0} g(y|\tilde{y}; x) = \lim_{y \to \infty} g(y|\tilde{y}; x) = \infty, \tag{54}$$

so that (51) always admits two positive solutions (or one double positive solution if $\tilde{y} = x$), one of the two being $y = \tilde{y}$. The other one is the solution of interest. However, it is not closed-form, except for specific values of β (see Table 3). More precisely, when $\beta = 1 - 1/d$ and d is an integer, the solution can be found by solving the following polynomial equation of degree d, for $z = y^{1/d}$:

$$(1-\beta) \sum_{\ell=1}^{d} \tilde{z}^{d-\ell} z^{\ell} - x = 0 \tag{55}$$

where $\tilde{z} = \tilde{y}^{1/d}$. Not surprisingly, the simplest case $\beta = 0$ ($d = 1$) leads us to $y = x$, and thus to $h_k^{\mathrm{ME}} = h_k^{\mathrm{H}}$. The case $\beta = 0.5$ ($d = 2$) is more interesting. The extraction of the positive root of (55) then provides the following update formula:

$$h_k^{\mathrm{ME}} = \frac{\tilde{h}_k}{4} \left(\sqrt{1 + 8\frac{h_k^{\mathrm{H}}}{\tilde{h}_k}} - 1 \right)^2. \tag{56}$$

Let us remark that this expression does not correspond to a multiplicative update, although it ensures that positivity is maintained.

[3] Indeed, we can prove that the reversed inequality holds for all $\beta < 0$, while no systematic result is known for $\beta > 2$.

$\beta \leq 0$	$0 \leq \beta \leq 1$	$1 \leq \beta \leq 2$	$\beta \geq 2$	d
0	0	2	2	1
-1	$1/2$	$3/2$	3	2
-2	$2/3$	$4/3$	4	3
-3	$3/4$	$5/4$	5	4

Table 3: Values of β for which ME updates are closed-form, by root extraction of polynomials of degree d.

Case 2 : $\beta \in (1, 2]$ In that case we have

$$g(y|\tilde{y}; x) = \frac{1}{\beta} y^{\beta} - \frac{1}{\beta - 1} x y^{\beta - 1} + cst. \tag{57}$$

$g(y|\tilde{y}; x)$ tends toward ∞ for $y \to \infty$, but it remains finite for $y \to 0$. As a consequence, Eq. (51) only admits the trivial solution $y = \tilde{y}$ if $g(\tilde{y}|\tilde{y}; x) > g(0|\tilde{y}; x)$, and also the unwanted solution 0 if $g(\tilde{y}|\tilde{y}; x) = g(0|\tilde{y}; x)$. It is only when $g(\tilde{y}|\tilde{y}; x) < g(0|\tilde{y}; x)$ that a positive, non trivial solution exists. This solution is closed-form for specific values of β given in Table 3. They correspond to $\beta = 1 + 1/d$, where d is an integer. Eq. (51) then amounts to solve the following polynomial equation of degree d, for $z = y^{1/d}$:

$$\sum_{\ell=0}^{d} \tilde{z}^{d-\ell} z^{\ell} - (d+1)\, x = 0, \tag{58}$$

with $\tilde{z} = \tilde{y}^{1/d}$. The simplest case is $\beta = 2$ ($d = 1$), and the solution is then given by $y = 2x - \tilde{y}$ if $\tilde{y} < 2x$, which yields the overrelaxed update

$$h_k^{\mathrm{ME}} = 2 h_k^{\mathrm{H}} - \tilde{h}_k, \tag{59}$$

provided that $\tilde{h}_k < 2 h_k^{\mathrm{H}}$. In the case $\beta = 1.5$ ($d = 2$), a positive ME update exists if $\tilde{h}_k < 3 h_k^{\mathrm{H}}$, and it takes the following form:

$$h_k^{\mathrm{ME}} = \frac{\tilde{h}_k}{4} \left(\sqrt{12 \frac{h_k^{\mathrm{H}}}{\tilde{h}_k} - 3} - 1 \right)^2. \tag{60}$$

As we need an update strategy that is defined everywhere, we propose to rely on a linear mixture between the MM update and a prolonged version of ME, defined as

$$h_k^{\theta} = \theta h_k^{\mathrm{pME}} + (1 - \theta) h_k^{\mathrm{MM}} \tag{61}$$

where $\theta \in (0, 1)$ and h_k^{pME} prolongs the ME update by zero when the latter does not exist:

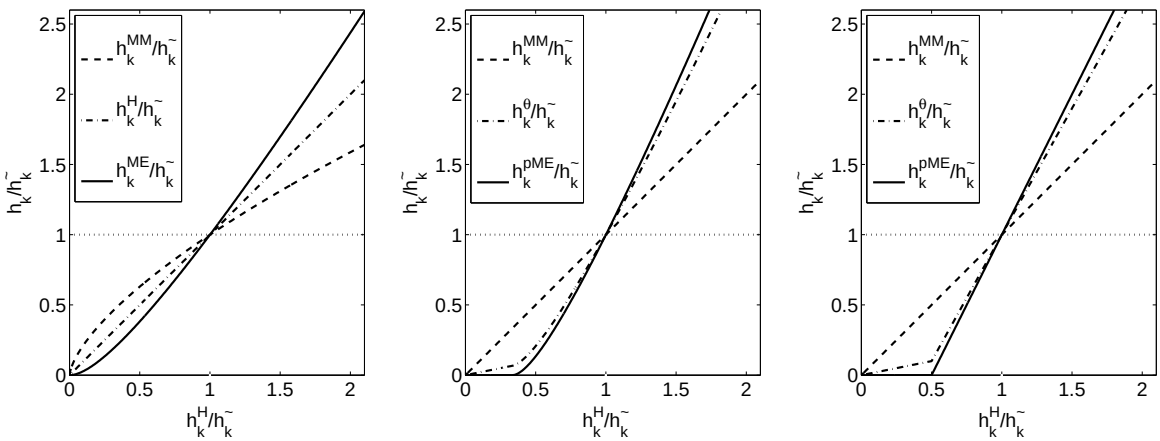
$$h_k^{\mathrm{pME}} = \begin{cases} h_k^{\mathrm{ME}} & \text{if } h_k^{\mathrm{ME}} \text{ is defined} \\ 0 & \text{otherwise} \end{cases} \tag{62}$$

See Figure 3 for an illustration. It is mathematically easy to check that h_k^{θ} fulfills Eq. (18) for all $\theta \in [0, 1]$, and that positivity is maintained for all $\theta \in [0, 1)$. In practice, values of θ near one may be favored to produce larger steps.

When $\beta < 0$ or $\beta > 2$, similar analyses can be conducted. In particular, there are specific values of β for which a closed-form expression of ME updates is available according to Table 3.

When $\beta < 0$, ME updates always exist since (53) and (54) still hold. Moreover, they provide non-increasing values of $C(\mathbf{h})$, while the latter monotonicity property is not yet proved for the heuristic algorithm. However, simulations tend to indicate that the heuristic algorithm is faster than the ME algorithm (which is itself faster than the MM algorithm) in the case $\beta < 0$. This is in conformity with the fact that ME steps can then be proved to be smaller than heuristic steps (on the basis of the reversed inequality mentioned in Footnote 3).

When $\beta > 2$, ME updates do not necessarily exist, akin to the case $\beta \in (1, 2]$. When they exist, they provide nonincreasing values of $C(\mathbf{h})$, while the latter is not yet proved for the heuristic formula. However, since this range of β values does not seem of practical interest, we will not go further into a detailed analysis here.



(a) $\beta = 0.5$ (b) $\beta = 1.5$ (b) $\beta = 2$

Figure 3: Normalized updates $h_k/\tilde{h}_k$ as functions of $h_k^{\mathrm{H}}/\tilde{h}_k$ ($\theta = 0.8$). The region between the dotted, horizontal line and the solid line correspond to the steps that fulfill Eq. (18). The larger departure from the horizontal line, the larger the step.

4.4 Implementation and complexity of the algorithms

The update rules of the algorithms presented above can all be expressed as functions of the ratio $h_k^{\mathrm{H}}/\tilde{h}_k = \nabla^-_{h_k} C(\tilde{\mathbf{h}})/\nabla^+_{h_k} C(\tilde{\mathbf{h}})$, which dominates the complexity of the algorithms. Fortunately, the latter ratio takes a simple matrix form that leads to efficient implementations. As such, getting back to the original factorization problem, the heuristic update for factors $\mathbf{H}$ and $\mathbf{W}$ can conveniently be expressed in the following matrix form

$$\mathbf{H} \leftarrow \mathbf{H}.\frac{\mathbf{W}^T\left[(\mathbf{WH})^{\cdot(\beta-2)}.\mathbf{V}\right]}{\mathbf{W}^T\left[\mathbf{WH}\right]^{\cdot(\beta-1)}} \tag{63}$$

$$\mathbf{W} \leftarrow \mathbf{W}.\frac{\left[(\mathbf{WH})^{\cdot(\beta-2)}.\mathbf{V}\right]\mathbf{H}^T}{\left[\mathbf{WH}\right]^{\cdot(\beta-1)}\mathbf{H}^T} \tag{64}$$

where the division $\cdot/\cdot$ is here taken entrywise. The simplicity of these update rules has undoubtedly contributed to the popularity of NMF. The MM update simply involves bringing the corrective ratio to the power $\gamma(\beta)$, and the ME update involves applying a function specific to the value of β. Hence, the algorithms have similar complexity $\mathcal{O}(FKN)$ and their implementation take simple forms. *MATLAB implementations of the proposed algorithms will be made available online at time of publication.*

5 Simulations

In this section we report performance results of β-NMF algorithms for the specific values $\beta = \{0.5, 1.5, 2\}$. These values are chosen for their practical interest and because a simple ME algorithm exists in their cases. As such this section will evidence the performance improvement brought by the ME approach over the MM or heuristic approaches, with similar computational burden. More precisely, the ME algorithm considered in this section is the mixture of prolonged ME and MM, defined by Eq. (61) and with $\theta = 0.95$, but we will still refer to it as ME for simplicity. The algorithms for all three considered values of β are compared on small-sized synthetic data in Section 5.1. The algorithms for $\beta = 0.5$ are analyzed in Section 5.2 on the basis of a small music transcription example as this specific value of β has proven efficient for this task (FitzGerald et al., 2009; Vincent et al., 2010).

In the following results we will display the cost values through iterations as well as, following (Gonzalez and Zhang, 2005), "KKT residuals". The residuals allow to monitor convergence to a stationary point and are here defined as

$$\mathrm{KKT}(\mathbf{W}) = \| \min\left\{\mathbf{W}, [(\mathbf{WH})^{\cdot(\beta-2)}.(\mathbf{WH}-\mathbf{V})]\mathbf{H}^T\right\} \|_1/FK \tag{65}$$

$$\mathrm{KKT}(\mathbf{H}) = \| \min\left\{\mathbf{H}, \mathbf{W}^T[(\mathbf{WH})^{\cdot(\beta-2)}.(\mathbf{WH}-\mathbf{V})]\right\} \|_1/KN. \tag{66}$$

They are meant to converge to zero, by Eq. (17). Again, the monotonicity of the heuristic, MM and ME algorithms does not imply convergence of the iterates to a stationary point. Hence, displaying the KKT residuals allows to experimentally check whether convergence is achieved in practice.

One iteration of each algorithm consists of updating $\mathbf{W}$ given $\mathbf{H}^{(i-1)}$ and $\mathbf{H}$ given $\mathbf{W}^{(i)}$, and then normalize $\mathbf{W}^{(i)}$ and $\mathbf{H}^{(i)}$ to eliminate trivial scale indeterminacies that leave the cost function unchanged. The normalization step consists of rescaling each column of $\mathbf{W}$ so that $\|\mathbf{w}_k\|_1 = 1$ and rescale the kth row of $\mathbf{H}$ accordingly. The normalization step is not required *per se* but is useful to display and compare the KKT residuals, which are scale-sensitive.

5.1 Factorization of synthetic data

We consider a synthetic data matrix $\mathbf{V}$ constructed as $\mathbf{V} = \mathbf{W}^*\mathbf{H}^*$ where the ground truth factors are generated as the absolute values of Gaussian noise.[4] The matrix can be exactly factorized so that all algorithms should converge to a solution such that $D(\mathbf{V}|\mathbf{WH}) = 0$. The dimensions are $F = 10$, $N = 25$, $K = 5$. The algorithms (heuristic, MM, ME for $\beta = 0.5$, MM and ME for $\beta = \{1.5, 2\}$) are run for 10^5 iterations and initialized with common positive random values. Fig. 4, 5 and 6 display for each of the 3 values of β the normalized cost values $D(\mathbf{V}|\mathbf{WH})/FN$, the KKT residuals, as well as "fit residuals" computed as $\|\mathbf{W}^{(i)} - \hat{\mathbf{W}}\|_F/FK$ and $\|\mathbf{H}^{(i)} - \hat{\mathbf{H}}\|_F/KN$, where $\hat{\mathbf{W}}$ and $\hat{\mathbf{H}}$ are the factor estimates at the end of the 10^5 iterations and $\|.\|_F$ is the Frobenius norm. The fit residuals allow to measure the closeness of the current iterates to their end value.

The cost values in all three cases converge to zero as an exact factorization is reached (oscillations appear in the end iterations as machine precision is reached). Convergence is achieved in all three cases, as shown by both the cost values and KKT residuals. The ME algorithm provides fastest convergence in every case (this was checked with other initializations as well). The fit residuals show that full convergence will not need be attained to obtain satisfying solutions for most applications as the fit residual will be considered sufficiently small after a few hundred iterations. Note that the factor iterates do not necessarily converge to the ground truth values $\mathbf{W}^*$ and $\mathbf{H}^*$ because of the identifiability ambiguities inherent to NMF (Donoho and Stodden, 2004; Laurberg et al., 2008).

The CPU time required by each algorithm for the 10^5 iterations is about 60s for $\beta = \{0.5, 1.5\}$ and 20s for $\beta = 2$, including the computation of the cost values and KKT residuals and run on Mac 2.6 GHz with 2 Go RAM in a MATLAB implementation. The ME algorithm is marginally more expensive than MM, itself only slightly more expensive than the heuristic algorithm, for $\beta = 0.5$. The CPU time needed with $\beta = 2$ is considerably lower thanks to simplifications in Eq. (63) and (64). Indeed, in latter case the term $(\mathbf{WH})\mathbf{H}^T$ appearing at the denominator can more efficiently be computed as $\mathbf{W}(\mathbf{HH}^T)$, which involves a multiplication of matrices with smaller size.

5.2 Factorization of real data

This section addresses the comparison of the heuristic, MM and ME algorithms for $\beta = 0.5$ applied to an audio spectrogram. We consider the short piano sequence of (Févotte et al., 2009), recorded in live conditions, composed of 4 musical notes, played all at once in the first measure and then played by pairs in all possible combinations in the subsequent measures. A magnitude spectrogram of the audio signal is computed, leading to nonnegative matrix data $\mathbf{V}$ of size $F = 513$ frequency bins by $N = 674$ time frames. The data is represented in Fig. 7.

As discussed in (Févotte et al., 2009), K was set to 6 so as to retrieve in $\mathbf{W}$ the individual spectra of each of the 4 notes and supplementary spectra corresponding to transients and residual noise. The three algorithms were initialized with common positive random values and run for 10^5 iterations. Figure 8 displays the cost values and KKT residuals along the 10^5 iterations. Other initializations yielded sensibly similar plots. It was manually checked that the algorithms converged to the desired "ground-truth" solution, i.e., the notes, transients and residual noise spectra are correctly unmingled. The three plots show that the ME provides fastest convergence overall though, judging from the KKT residuals, it appears that convergence is not achieved within the 10^5 iterations. However, the musical pitch values (computed from $\mathbf{W}$ at every iteration) converge to their ground truth values after only 30, 50 and 580 iterations for ME, heuristic and MM, respectively. Note that in some other runs from other random initializations the pitch values converged faster with the heuristic algorithm than with ME, and it was found that MM is generally slower than the two other algorithms.

Let us mention that besides this music transcription example we run simulations using the MM and ME algorithms with $\beta \in \{1, 2\}$ on the CBCL face dataset (CBCL, 2000). This dataset has often been considered in the NMF literature, in particular in Lee & Seung's seminal paper (Lee and Seung, 1999). The results in terms of convergence behavior were very similar to those of Figure 8 and are not reproduced

[4] E.g., in MATLAB notation `V = abs(randn(F,K))*abs(randn(K,N))` .

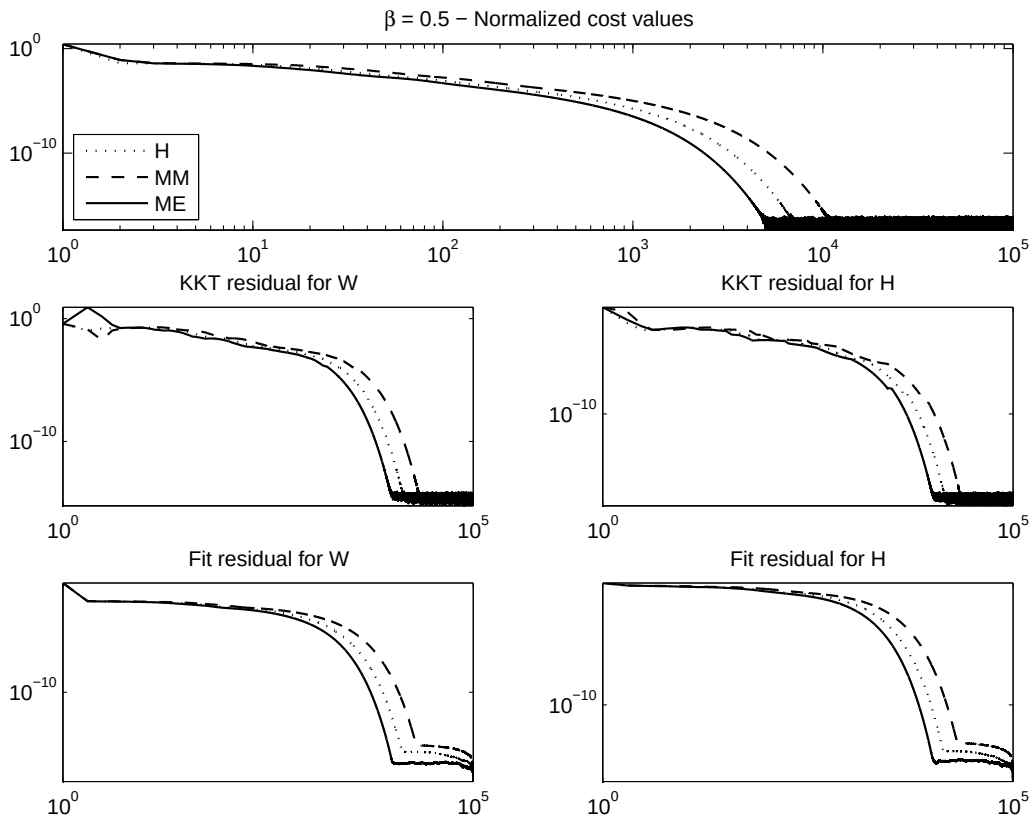


Figure 4: One run of the heuristic (H), ME and MM algorithms on synthetic data with $\beta = 0.5$. Logarithmic scales for both x- and y- axes.

here : the MM algorithm produces a sharper decrease of the cost function in the first few iterations but is then outperformed by the ME algorithm which in the end converge faster. Convergence of the KKT residuals within the 10^5 iterations is not clear, but the learnt dictionaries stabilizes (visually) after a few hundred iterations.

6 Variants of β-NMF

In this section we briefly discuss how some common variants of NMF, penalized NMF and convex-NMF, can be handled under NMF with the β-divergence.

Penalized β-NMF Supplementary functions of $\mathbf{W}$ and/or $\mathbf{H}$ are often added to the cost function (3) so as to induce some sort of regularization of the factor estimates or so as to reflect prior belief (e.g., in Bayesian MAP estimation). When such penalty terms are separable in the columns of $\mathbf{H}$ or in the rows of $\mathbf{W}$, penalized NMF essentially amounts to solving the following optimization problem :

$$\min_{\mathbf{h}} C_P(\mathbf{h}) \stackrel{\text{def}}{=} D(\mathbf{v}|\mathbf{W}\mathbf{h}) + L(\mathbf{h}) \text{ subject to } \mathbf{h} \geq 0 \tag{67}$$

where $L(\mathbf{h})$ is the penalty term. An auxiliary function to $C_P(\mathbf{h})$ is readily given by

$$G_P(\mathbf{h}|\tilde{\mathbf{h}}) \stackrel{\text{def}}{=} G(\mathbf{h}|\tilde{\mathbf{h}}) + L(\mathbf{h}) \tag{68}$$

where $G(\mathbf{h}|\tilde{\mathbf{h}})$ is any auxiliary function to $C(\mathbf{h}) = D(\mathbf{v}|\mathbf{W}\mathbf{h})$. MM or ME algorithms can then be designed on a case-by-case basis. Let us consider a short example for illustration : ℓ_1-norm regularization, i.e.,

$$L(\mathbf{h}) = \lambda \sum_k h_k. \tag{69}$$

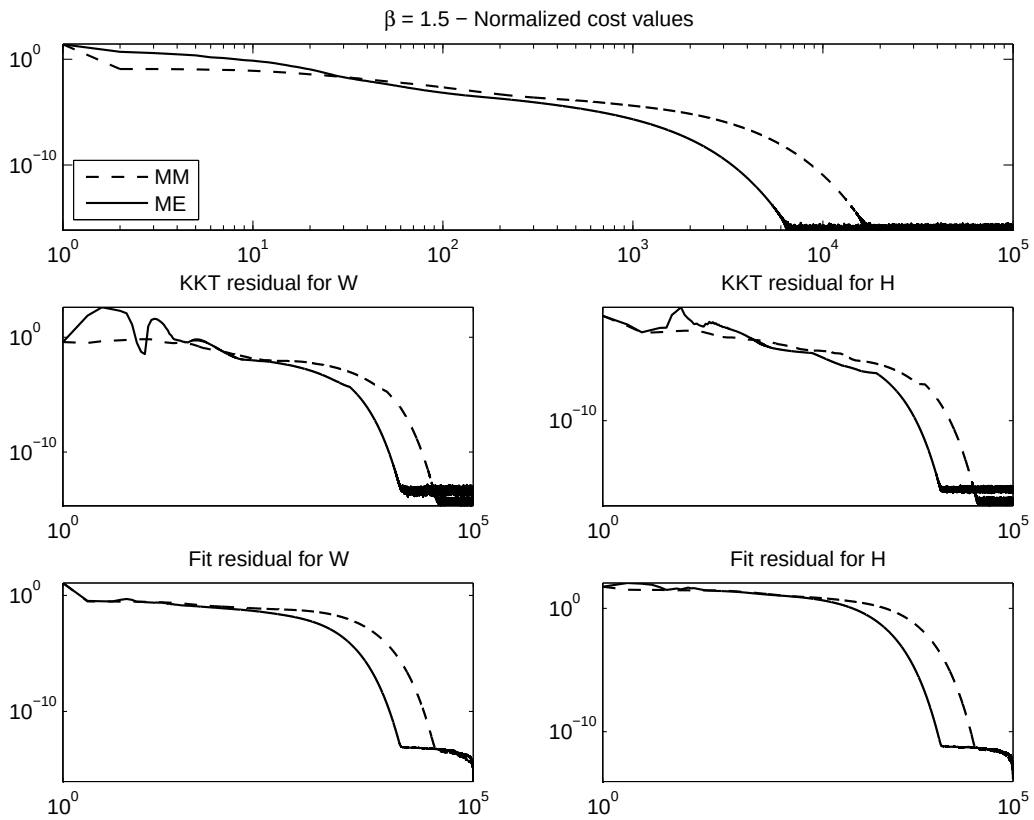


Figure 5: One run of the ME and MM algorithms on synthetic data with $\beta = 1.5$. Logarithmic scales for both x- and y- axes.

where λ is a positive weight parameter. In that case, using the separable auxiliary function designed in Section 3.2 and Eq. (36), the gradient of the penalized auxiliary function writes

$$\nabla_{h_k} G_L(\mathbf{h}|\tilde{\mathbf{h}}) = \sum_f w_{fk} \left[\breve{d}'\left(v_f | \tilde{v}_f \frac{h_k}{\tilde{h}_k} \right) + \widehat{d}'(v_f|\tilde{v}_f) \right] + \lambda.$$

The MM algorithm for ℓ_1 penalized β-NMF takes a very simple form for $\beta \leq 1$, such that

$$h_k = \tilde{h}_k \left(\frac{\sum_f w_{fk}\, v_f\, \tilde{v}_f^{\beta-2}}{\sum_f w_{fk}\, \tilde{v}_f^{\beta-1} + \lambda} \right)^{\gamma(\beta)}. \tag{70}$$

This in particular leads to ℓ_1-regularized NMF algorithms for KL-NMF and IS-NMF with proven monotonicity. An update similar to Eq. (70) is obtained for $\beta \geq 2$ but the λ term appears through its sign opposite at the numerator, instead of appearing at the denominator. Hence the nonnegativity constraint may become active and must be treated carefully; in that case our result coincides with similar findings in (Pauca et al., 2006; Mørup and Clemmensen, 2007) for the specific case of ℓ_1-regularized NMF with the Euclidean distance ($\beta = 2$). In the case $\beta \in (1, 2)$ the MM algorithm does not come up with a simple closed-form update, which supports the fact in the penalized case handy algorithms may only come on a case-by-case basis. This is similar to Expectation-Maximization (EM) procedures for MAP estimation, in which the E-step is essentially unchanged but where the M-step might become intractable because of the penalty term. ME algorithms can also be designed for the ℓ_1-regularized problem and as a matter of fact it can be shown that the results of Table 3 (i.e., the values of β for which a closed-form update exists) still hold in that case.

Convex β-NMF In some recent NMF-related works the dictionary $\mathbf{W}$ is constrained to belong to a known subspace $\mathbf{S} \in \mathbb{R}_+^{F \times M}$ such that

$$\mathbf{W} = \mathbf{S}\mathbf{L} \tag{71}$$

where $\mathbf{L} \in \mathbb{R}_+^{M \times K}$. For example in (Ding et al., 2010) the columns of $\mathbf{W}$ are assumed to be linear combinations (with unknown expansion coefficients) of data points (columns of $\mathbf{V}$), so as to enforce the

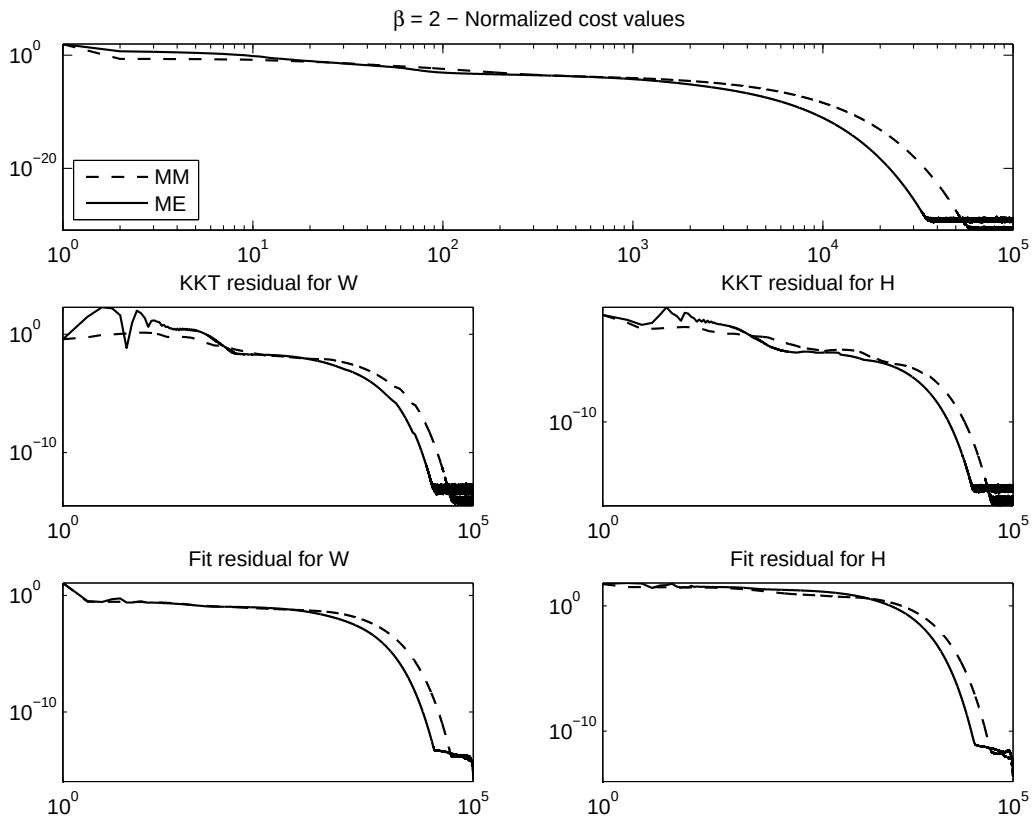


Figure 6: One run of the ME and MM algorithms on synthetic data with $\beta = 2$. Logarithmic scales for both x- and y- axes.

dictionary to be composed of *data centroids*, while in (Vincent et al., 2010) the dictionary elements are assumed to be linear combinations of narrow band spectra, so as to enforce harmonicity and smoothness of the dictionary. The term "convex-NMF" was introduced by Ding et al. (2010) to express the idea that $\mathbf{W}$ belongs to the convex set made of all nonnegative linear combinations of elements of $\mathbf{S}$, but this does not make the optimization problem convex in itself, in the general case.

In this setting, the dictionary update is tantamount to solving

$$\min_{\mathbf{L}} C_{cv}(\mathbf{L}) \stackrel{\text{def}}{=} D(\mathbf{V}|\mathbf{SLH}) = \sum_{fn} d\left(v_{fn}|\sum_{mk} s_{fm}h_{kn}l_{mk}\right) \quad \text{subject to } \mathbf{L} \geq 0. \tag{72}$$

As a matter of fact, we will show that this matricial optimization problem can be turn into vectorial nonnegative linear regression so that the results of Section 4 holds. Given some mappings $(f,n) \in \{1,F\} \times \{1,N\} \to p \in \{1,FN\}$ and $(m,k) \in \{1,M\} \times \{1,K\} \to q \in \{1,MK\}$ let us introduce the following variables : $\mathbf{T}$ is the matrix of dimension $FN \times MK$ with coefficients $t_{pq} = s_{fm}h_{kn}$, $\underline{\mathbf{v}}$ is the column vector of size FN with coefficients $\underline{v}_p = v_{fn}$, $\underline{\mathbf{l}}$ is the column vector of size MK with coefficients $\underline{l}_q = l_{mk}$. Then we have

$$D(\mathbf{V}|\mathbf{SLH}) = \sum_{p} d\left(\underline{v}_p|\sum_{q} t_{pq}\,\underline{l}_q\right) \tag{73}$$

and thus the estimation of $\mathbf{L}$ amounts to the approximation $\underline{\mathbf{v}} \approx \mathbf{T}\,\underline{\mathbf{l}}$. As such, any of the algorithms described in Section 4 can be employed for this task. As before, the resulting vectorial updates can be turned into matricial updates, leading to simple and efficient implementations. For example, the MM update reads

$$\mathbf{L} \leftarrow \mathbf{L}.\left(\frac{\mathbf{S}^T\left[(\mathbf{SLH})^{\cdot(\beta-2)}.\mathbf{V}\right]\mathbf{H}^T}{\mathbf{S}^T\left[(\mathbf{SLH})^{\cdot(\beta-1)}\right]\mathbf{H}^T}\right)^{\cdot\gamma(\beta)}. \tag{74}$$

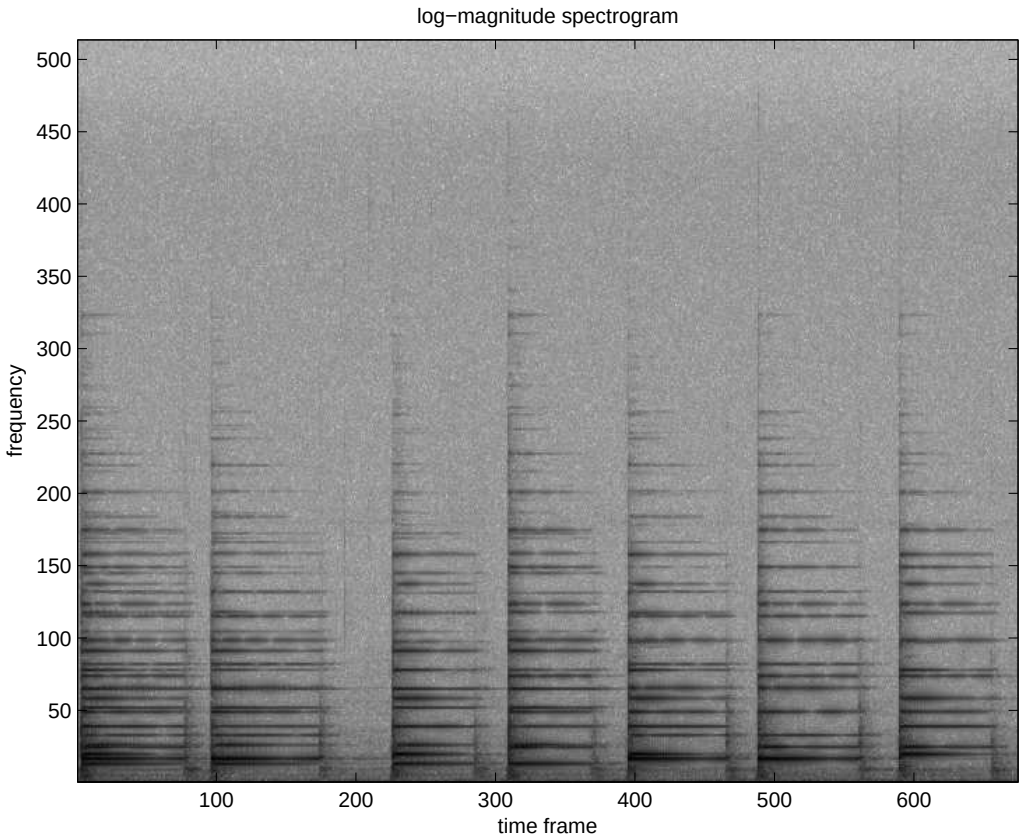


Figure 7: Log-magnitude spectrogram $\log \mathbf{V}$ of a short piano sequence.

This result proves the monotonicity of some of the algorithms derived heuristically in (Vincent et al., 2010) and also extends the results of (Ding et al., 2010) for convex NMF with the Euclidean distance to the more general β-divergence.[5]

7 Conclusions

This paper has addressed NMF with the β-divergence. The problem may be reduced to a mere nonnegative linear regression problem and our approach is based on the construction of an auxiliary function $G(\mathbf{h}|\tilde{\mathbf{h}})$ which majorizes the objection function $C(\mathbf{h})$ everywhere and is tight for $\mathbf{h} = \tilde{\mathbf{h}}$. The auxiliary function unifies existing auxiliary functions for the Euclidean distance and the KL divergence (Lee and Seung, 2001), for the "generalized divergence" of (Kompass, 2007) (in essence the β-divergence on its convex part, i.e., $\beta \in [1, 2]$) and for the IS divergence (Cao et al., 1999). Various descent algorithms may then be derived from this auxiliary function. As such, the findings of this paper may be summarized as follows.

- The MM algorithm based on the described auxiliary function is shown to yield multiplicative algorithms for $\beta \in \mathbb{R}$, as described by Eq. (38) and independenty obtained by Nakano et al. (2010). For $\beta \in [1, 2]$ (interval of values for which the β-divergence is convex), the MM algorithm coincides with the heuristic algorithm given by Eq. (43), as already known from Kompass (2007).
- In Section 4.2, we prove the monotonicity of the heuristic algorithm for $\beta \in (0, 1)$ by proving the inequality $G(\mathbf{h}^{\mathrm{H}}|\tilde{\mathbf{h}}) \leq G(\tilde{\mathbf{h}}|\tilde{\mathbf{h}})$. Hence, aggregating the existing monotonicity results for $\beta = 0$ and $\beta \in [1, 2]$, it can now be claimed that the heuristic algorithm is monotone for $\beta \in [0, 2]$, which is the range of values of practical interest that has been considered in the literature.
- In Section 4.3, we introduced the concept of maximization-equalization (ME) algorithms. Such algorithms are exhibited for specific values of β, in particular for $\beta \in \{0, 0.5, 1.5, 2\}$ which are

[5] More precisely, Ding et al. (2010) consider a "semi"-NMF version where $\mathbf{S} = \mathbf{V}$ and the data is allowed to be real-valued while the nonnegativity constraint is solely imposed on $\mathbf{L}$ and $\mathbf{H}$; our results do not apply to this more general framework but only to the special case where $\mathbf{V}$ in nonnegative.

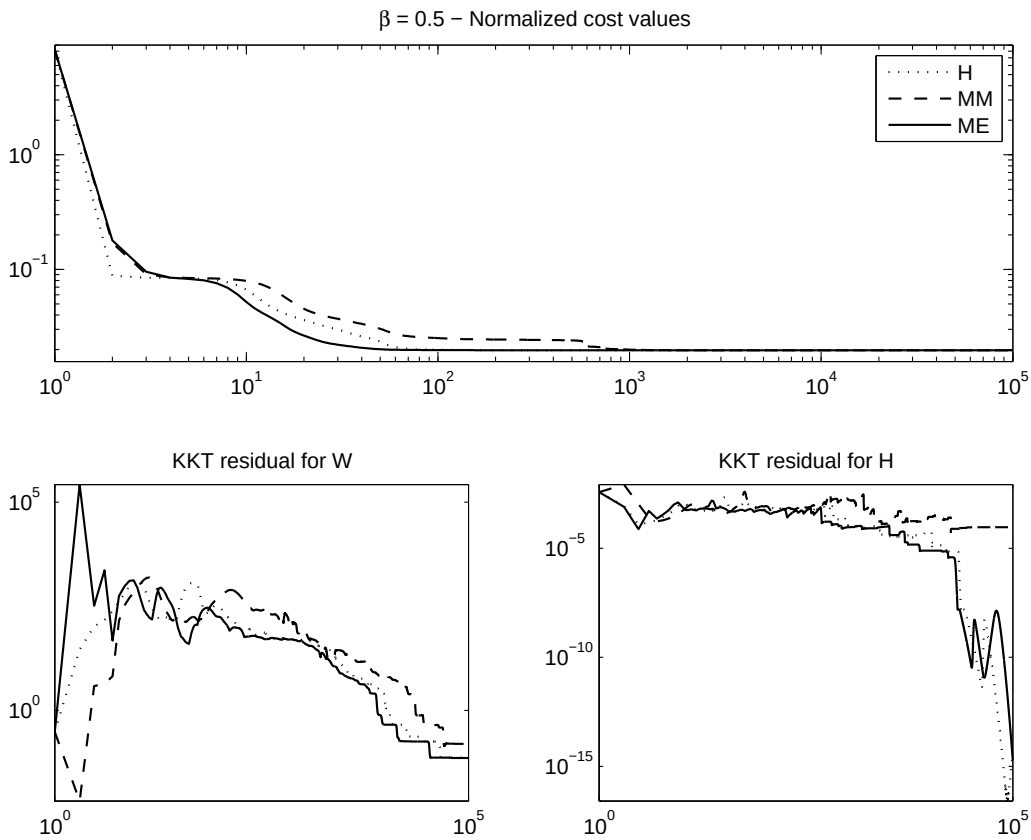


Figure 8: One run of the heuristic (H), MM and ME algorithms on the piano magnitude spectrogram with $\beta = 0.5$. Logarithmic scales for both x- and y- axes.

values of practical interest. For $\beta = 0$ (IS divergence) the ME algorithm coincides with the heuristic algorithm, whose monotonicity already holds from (Cao et al., 1999). For other values of β the ME algorithms are nonmultiplicative. For $\beta \in \{0.5, 1.5, 2\}$ they amount to solving polynomial equations of order 1 or 2. The result section has illustrated the faster convergence of the ME approach w.r.t to MM or heuristic, with equivalent complexity.

- Finally, in Section 6 we have considered variants of NMF with the β-divergence. We have explained how penalty terms may be handled in the auxiliary function setting; in particular we have presented simple multiplicative algorithms for ℓ_1 regularized KL or IS NMF. Then, we have shown how the algorithms constructed for plain NMF holds for convex-NMF, generalizing and proving the monotonicity of existing algorithms.

As for perspectives, the present work leaves two important questions unanswered. The first one is the monotonicity of the heuristic algorithm for $\beta \notin [0, 2]$. The monotonicity is observed in practice but we have not been able to come up with proofs in the presented setting. Either other approaches need to be followed or a different type of auxiliary functions than the one presented here needs to be envisaged. As suggested in Section 2.1, the convex-concave decomposition of the β-divergence is not unique and other decomposition than the "natural" one employed in this paper may lead to auxiliary functions that more closely fit to the criterion. The second, probably more ambitious question is the convergence of the sequence of iterates produced by the proposed algorithms a stationary point. Partial results exist for Euclidean NMF (Lin, 2007), convergence of multiplicative rules for nonnegative linear regression has been studied in a few cases, see, e.g., (Titterington, 1987; De Pierro, 1993; Eggermont and LaRiccia, 1998), but general results for the β-divergence are still lacking.

References

Basu, A., Harris, I. R., Hjort, N. L., and Jones, M. C. (1998). Robust and efficient estimation by minimising a density power divergence robust and efficient estimation by minimising a density power divergence. *Biometrika*, 85(3):549–559.

Berry, M. W., Browne, M., Langville, A. N., Pauca, V. P., and Plemmons, R. J. (2007). Algorithms and applications for approximate nonnegative matrix factorization. *Computational Statistics & Data Analysis*, 52(1):155–173.

Bertin, N., Févotte, C., and Badeau, R. (2009). A tempering approach for Itakura-Saito non-negative matrix factorization. With application to music transcription. In *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP'09)*, pages 1545 –1548, Taipei, Taiwan.

Brunet, J.-P., Tamayo, P., Golub, T. R., and Mesirov, J. P. (2004). Metagenes and molecular pattern discovery using matrix factorization. In *Proceedings of the National Academy of Sciences*, pages 4164–4169.

Cao, Y., Eggermont, P. P. B., and Terebey, S. (1999). Cross Burg entropy maximization and its application to ringing suppression in image reconstruction. *IEEE Transactions on Image Processing*, 8(2):286–292.

CBCL (2000). Face database #1. `http://cbcl.mit.edu/software-datasets/FaceData2.html`. Center for Biological and Computational Learning at MIT.

Cichocki, A. and Amari, S. (2010). Families of Alpha- Beta- and Gamma- divergences: Flexible and robust measures of similarities. *Entropy*, 12(6):1532–1568.

Cichocki, A., Zdunek, R., and Amari, S. (2006). Csiszar's divergences for non-negative matrix factorization: Family of new algorithms. In *Proc. 6th International Conference on Independent Component Analysis and Blind Signal Separation (ICA'06)*, pages 32–39, Charleston SC, USA.

Daube-Witherspoon, M. and Muehllehner, G. (1986). An iterative image space reconstruction algorthm suitable for volume ECT. *IEEE Transactions on Medical Imaging*, 5(5):61 – 66.

De Pierro, A. R. (1993). On the relation between the ISRA and the EM algorithm for positron emission tomography. *IEEE Trans. Medical Imaging*, 12(2):328–333.

Dessein, A., Cont, A., and Lemaitre, G. (2010). Real-time polyphonic music transcription with non-negative matrix factorization and beta-divergence. In *Proc. 11th International Society for Music Information Retrieval Conference (ISMIR'2010)*.

Dhillon, I. S. and Sra, S. (2005). Generalized nonnegative matrix approximations with Bregman divergences. *Advances in Neural Information Processing Systems (NIPS)*, 19.

Ding, C. H. Q., Li, T., and Jordan, M. I. (2010). Convex and semi-nonnegative matrix factorizations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(1):45 – 55.

Donoho, D. and Stodden, V. (2004). When does non-negative matrix factorization give a correct decomposition into parts? In Thrun, S., Saul, L., and Schölkopf, B., editors, *Advances in Neural Information Processing Systems 16*. MIT Press, Cambridge, MA.

Drakakis, K., Rickard, S., de Frein, R., and Cichocki, A. (2007). Analysis of financial data using non-negative matrix factorization. *International Journal of Mathematical Sciences*, 6(2).

Eggermont, P. P. B. and LaRiccia, V. N. (1998). On EM-like algorithms for minimum distance estimation. http://www.udel.edu/FREC/eggermont/Preprints/emlike.pdf.

Eguchi, S. and Kano, Y. (2001). Robustifying maximum likelihood estimation. Technical report, Institute of Statistical Mathematics. Research Memo. 802.

Févotte, C., Bertin, N., and Durrieu, J.-L. (2009). Nonnegative matrix factorization with the Itakura-Saito divergence. With application to music analysis. *Neural Computation*, 21(3):793–830.

Févotte, C. and Cemgil, A. T. (2009). Nonnegative matrix factorisations as probabilistic inference in composite models. In *Proc. 17th European Signal Processing Conference (EUSIPCO'09)*, pages 1913–1917, Glasgow, Scotland.

FitzGerald, D., Cranitch, M., and Coyle, E. (2009). On the use of the beta divergence for musical source separation. In *Proc. Irish Signals and Systems Conference*.

Gonzalez, E. F. and Zhang, Y. (2005). Accelerating the Lee-Seung algorithm for non-negative matrix factorization. Technical report, Rice University.

Greene, D., Cagney, G., Krogan, N., and Cunningham, P. (2008). Ensemble non-negative matrix factorization methods for clustering protein-protein interactions. *Bioinformatics*, 24(15):1722–1728.

Hunter, D. and Lange, K. (2004). A tutorial on MM algorithms. *The American Statistician*, 58:30 – 37.

Kompass, R. (2007). A generalized divergence measure fon nonnegative matrix factorization. *Neural Computation*, 19(3):780–791.

Laurberg, H., Christensen, M. G., Plumbley, M. D., Hansen, L. K., and Jensen, S. H. (2008). Theorems on positive data: On the uniqueness of nmf. *Computational Intelligence and Neuroscience*, Article ID 764206.

Lee, D. D. and Seung, H. S. (1999). Learning the parts of objects with nonnegative matrix factorization. *Nature*, 401:788–791.

Lee, D. D. and Seung, H. S. (2001). Algorithms for non-negative matrix factorization. In *Advances in Neural and Information Processing Systems 13*, pages 556–562.

Lin, C.-J. (2007). On the convergence of multiplicative update algorithms for nonnegative matrix factorization. *IEEE Transactions on Neural Networks*, 18:1589–1596.

Lucy, L. B. (1974). An iterative technique for the rectification of observed distributions. *Astronomical Journal*, 79:745–754.

Mørup, M. and Clemmensen, L. H. (2007). Multiplicative updates for the LASSO. In *Proc. IEEE Workshop on Machine Learning for Signal Processing (MLSP'07)*.

Nakano, M., Kameoka, H., Le Roux, J., Kitano, Y., Ono, N., and Sagayama, S. (2010). Convergence-guaranteed multiplicative algorithms for non-negative matrix factorization with beta-divergence. In *Proc. IEEE International Workshop on Machine Learning for Signal Processing (MLSP'2010)*.

O'Grady, P. D. (2007). *Sparse separation of under-determined speech mixtures*. PhD thesis, National University of Ireland Maynooth.

O'Grady, P. D. and Pearlmutter, B. A. (2008). Discovering speech phones using convolutive non-negative matrix factorisation with a sparseness constraint. *Neurocomputing*, 72(1-3):88 – 101.

Pauca, V. P., Piper, J., and Plemmons, R. J. (2006). Nonnegative matrix factorization for spectral data analysis. *Linear Algebra and its Applications*, 416:29–47.

Richardson, W. H. (1972). Bayesian-based iterative method of image restoration. *Journal of the Optical Society of America*, 62:55–59.

Smaragdis, P. and Brown, J. C. (2003). Non-negative matrix factorization for polyphonic music transcription. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA'03)*.

Titterington, D. M. (1987). On the iterative image space reconstruction algorthm for ECT. *IEEE Trans. Medical Imaging*, 6(1):52–56.

Vincent, E., Bertin, N., and Badeau, R. (2010). Adaptive harmonic spectral decomposition for multiple pitch estimation. *IEEE Trans. Audio, Speech and Language Processing*, 18:528 – 537.

Xu, W., Liu, X., and Gong, Y. (2003). Document clustering based on non-negative matrix factorization. In *Proc. 26th International ACM SIGIR conference on research and development in information retrieval (SIGIR'03)*, pages 267–273, Toronto, Canada.